Building a corpus for identifying AI-generated vs human-written texts

With the increased prevalence of AI-generated content across domains, distinguishing human-written text from AI-generated text is becoming an important research challenge. We hope our project to construct a corpus that includes a variety of text types, can help analyze patterns and differences between AI and human generated text. We believe a corpus like this could contribute to AI-generated content detection, misinformation research, and the overarching AI ethics environment.

Source of the data

Our corpus will consist of human written and AI-generated texts sourced from multiple domains, such as movies and book reviews, to mirror more colloquial opinion pieces, scraped from platforms such as Letterboxd or goodreads, and news articles, to mirror more journalistic, professional text, sourced from open-access news websites such as Reutrers. For AI-generated text, we will use language models such as GPT-4, Claude, and Gemini to generate content that mimics the style and topics of the human-written samples in our dataset.

Type of text

The corpus will focus exclusively on English-language pieces, where opinion pieces will be mostly movie reviews such as https://letterboxd.com/film/oppenheimer-2023/, and news article will be scraped from sites such as https://www.reuters.com/legal/government/. The diversity of genres may change depending on feasibility; in which case we may focus specifically on one type of text and its AI-generated counterpart. We aim to obtain articles ranging from 300 to 1000 words and will either be written by professional journalists, reviewing platform users, or generated by AI models.

Corpus Size & Structure

We plan to collect approximately 500 human-written documents and generate 500 AI-written articles, resulting in a corpus of approximately one million words. Each document will be stored as an individual entry, accompanied by metadata such as the source (review, news, etc.), publication date, word count, topic category, and whether it is AI-generated or human-written. We do not plan to include discussion threads, additional comments, interactions, etc.

Annotation Plan

Given our limited time and team size, we will focus on a single high-value annotation strategy. We have below three approaches that we have considered, from which we will select one based on feasibility and impact (we've bolded what we believe is the most feasible for this project):

- *Document-Level Factuality & Misleading Language*: Instead of verifying each claim individually, we would assess the overall credibility of an article using a factuality rating scale (1 = would be completely false, 5 = fully factual). We could also mark whether the article uses misleading tactics such as exaggeration, emotional manipulation, lack of sources, or logical fallacies, although this may be time-consuming and difficult to operationalize.

- ***Source Attribution & Citation Behavio*r: alternative, we plan on annotating whether news articles include citations, whether those citations are real, and whether they reference authoritative sources. This approach will help assess the reliability of AI-generated news, and is quite feasible at a document-level.**
- *Persuasive Framing & Emotional Appeal*: We could alternatively categorize the dominant rhetorical techniques used in each article/piece, identifying whether the primary persuasive strategy is logical argumentation (logos), emotional appeal (pathos), authority-based persuasion (ethos), or sensationalism. This would allow us to compare how AI and human writers may structure persuasive narratives differently, and also allows us to use this type of annotation across news articles and opinion pieces alike.

To ensure consistency, we will create a clear annotation guide, that would be revised by one of the course instructors, and cross-check a portion of the annotations for inter-annotator agreement.

Storage Format

To ensure ease of access and processing, the corpus will be stored in JSON format, which would allow structured metadata storage and straightforward integration with NLP tools.

Potential Value

We believe a corpus like ours has the potential to be used in multiple research domains, including training models to detect AI-generated misinformation, assisting fact-checkers and news organizations in verifying sources, detecting AI-generated content across different domains, and developing tools that may help content moderation platforms distinguish between authentic human engagement and AI-generated posts. We hope this corpus contributes to ongoing conversations around AI ethics in news generation, and the broader implications of AI-generated content.